Risks and Control Measures for Building Trustworthy Autonomous Weapon Systems

Clara Maathuis[1] and Kasper Cools[2, 3]
[1]Open University of the Netherlands, Heerlen, The Netherlands
[2]Belgian Royal Military Academy, Brussel, Belgium
[3]Vrije Universiteit Brussel, Belgium

clara.maathuis@ou.nl
kasper.cools@mil.be

Abstract: This research examines the risks and control measures associated with building trustworthy Autonomous Weapon Systems (AWS), a rapidly evolving technology with various implications for military operations and international security. While AWS present advantages in precision and efficiency, they also imply operational, technical, and ethical challenges. Through a comprehensive analysis of relevant studies, this article identifies key risks inherent in AWS development, including algorithmic biases, unintended engagements, and cyber security vulnerabilities. For these, control measures are proposed to mitigate and avoid them, such as advanced fail-safe mechanisms, multi-layered human oversight protocols, and robust cyber security solutions. Particular attention is given to the role of meaningful human control as a fundamental mechanism for enhancing AWS trustworthiness without compromising operational effectiveness. The findings highlight the need for a dynamic, proactive, multidisciplinary risk-based approach to AWS development as trustworthy systems, emphasising the importance of international collaboration in establishing standardised risk assessment methodologies, trustworthiness benchmarks, and certification processes. Moreover, by systematically analysing both risks and control measures, this research provides a design framework for addressing the complex challenges of building trustworthy AWS in the context of evolving warfare technologies.

Keywords: Trustworthy AI, Autonomous weapons systems, Trustworthy AWS, Artificial intelligence, Military operations

1. Introduction

> *"The world of the future will be an ever more demanding struggle against the limitations of our intelligence, not a comfortable hammock in which we can lie down to be waited upon by our robot slaves."*
>
> *(Norbert Wiener)*

Recent advancements in the Artificial Intelligence (AI) domain showed an important transformation power in various societal domains. This allows the enhancement of existing operational capabilities and the development of new ones across various aspects such as surveillance, decision-making, and combat efficiency. These technologies enable faster data processing, improved situational awareness, automated threat detection, and more precise execution of complex missions. These advancements have also led to the development of Autonomous Weapon Systems (AWS), which can engage targets with minimal or no human intervention (Théron and Kott, 2019; Horowitz, 2021). Concurrently, both major powers and small countries/groups are investing in these technologies bringing them to a high place in the global military competition (Haner and Garcia, 2019; Boyles, 2021; Rosendorf, Smetana and Vranka, 2024). While offering significant strategic advantages, from enhancing operational efficiency to minimising human risk, AWS exemplify the convergence of AI and autonomous technologies in warfare, offering the potential for greater tactical efficiency while raising important concerns about human oversight, responsibility, and bias.

Various studies emphasise that the framing of AWS as "killer robots" oversimplifies the complexities of human-machine interaction in warfare, particularly in terms of moral judgement and decision-making. This debate highlights the importance of maintaining "meaningful human control" over such systems, which is seen as crucial for addressing the psychological and ethical dimensions of warfare that machines may not be able to replicate adequately (Johnson, 2022). At the same time, AWS vary significantly in their degrees of autonomy, ranging from planning-autonomous systems that can execute predefined tasks to more advanced learning-autonomous systems that adapt to dynamic environments. Planning systems are already widely used in precision guided weapons, but there is a growing unease about learning-autonomous systems, which can generalise from past experiences and adapt in unpredictable ways. The distinction between these levels of autonomy is crucial, as trust in AWS hinges on their predictability and reliability, particularly in complex combat scenarios. Trust is a dynamic process that involves cognitive and emotional factors, and it evolves through operators' interaction over time. In situations of uncertainty and risk, trust becomes even more

critical, as it underpins the human-AI collaboration needed to ensure mission success (Roff and Danks, 2018a; Mayer, 2023a; Cools and Maathuis, 2024). Additionally, human-to-robot trust differs fundamentally from interpersonal trust, with several dimensions shaping how humans interact with machines, including cognition-based trust, affect-based trust, and initial trustworthiness. These factors evolve as operators learn more about the system's behaviour and capabilities. This uncertainty, which arises from both the technical performance of AWS and their ability to adapt to changing conditions, makes trust dynamic and essential for effective human-machine collaboration. The willingness to rely on AWS in uncertain combat scenarios involves not only confidence in their capabilities, but also a readiness to accept vulnerability, as misplaced trust can have severe consequences. Consequently, developing AWS requires rigorous attention to predictability, reliability, and the reduction of operational complexity to ensure that trust can be established and maintained under unpredictable conditions (Mancini *et al.*, 2019; Warren and Hillas, 2020; Firlej and Taeihagh, 2021; Mayer, 2023).

While different strategic, practitioner, and academic studies are dedicated to understanding various technical, ethical, and legal dimensions characterising various life cycle phases of AWS, there is a need for an analytical approach to understand the dimensions that surround the building process of AWS as trustworthy systems in relation to potential risks and corresponding control measures that could be applied to them. Accordingly, the goal of this research is to critically analyse the risks and corresponding control measures surrounding the development of AWS in military operations. Therefore, the following research questions (RQ) are formulated:

RQ1: What are the risks occurring when building AWS in military operations?

RQ2: What are the control measures that could be applied to mitigate or avoid the risks of AWS in military operations?

To answer these research questions, a systematic literature review stance is considered in alignment with the PRISMA research methodology (Denyer and Tranfield, 2009; Page *et al.*, 2021). Through the findings of this study, this research intends to contribute to existing academic and practitioner efforts that bring awareness and propose concrete solutions for building and deploying safe, responsible, and trustworthy AWS.

The outline of the article is structured as follows. Section 2 discusses relevant studies. Section 3 presents the methodological approach considered to achieve the goals defined. Section 4 provides the taxonomy of risks identified. Section 5 discusses control measures for mitigating or avoiding the risks identified. At the end, concluding remarks and two directions for future research are provided and analysed.

2. Related Research

AWS offer significant benefits but also pose significant risks, as addressed by Pedron and da Cruz (2020). They emphasise that AWS could reduce human deployments in high-risk areas, thereby lowering opposition to military interventions. For instance, potential risks for AWS include dual-use technology repurposing, affordability raising deployment concerns, and accountability gaps due to legal ambiguity and technological limitations, complicating regulation. Moreover, machine learning systems struggle with interpretability, reliability in unseen environments, and decision-making delays. They are vulnerable to adversarial perturbations, and the use of facial recognition for targeting raises potential ethical issues. These technical risks reflect the need for robust cybersecurity and international regulation to manage AWS proliferation and use (Longpre, Storm and Shah, 2022; Cools and Maathuis, 2024).

The responsible use of AI in military systems is critical, as discussed in Schraagen (2023). Emphasising the importance of keeping human operators in the loop, current research highlights the need for explainability and transparency in AI (Zajac, 2023). Moreover, the analysis of risk and safety in large-scale socio-technological military systems is crucial, as discussed by Bakx and Nyce (2017). Evaluations that consider various system levels and their interconnectedness are extensively covered in the social sciences and Science, Technology, and Society (STS) literature. STS focuses on the connections between science, technology, and society, providing a multifaceted perspective. Dekker's "drift into failure" (Dekker, 2011) illustrates how complex system interactions can result in failures. Similarly, Vaughan's examination of the Challenger disaster (1996) shows the interplay of micro and macro-level factors contributing to risk. Despite these insights in high-stakes safety domains, many military risk and safety reports lack the analytical rigour needed to fully understand these systems. Therefore, combining empirical accounts with an STS perspective can provide a more comprehensive understanding of how risk and safety emerge from the interplay of social and technological factors.

Finally, the dynamics of Human-Machine Teaming introduce significant considerations for autonomous (weapon) systems, as discussed in (Akiyoshi, 2022). The rapid growth of autonomous HMT systems (A-HMT-S) capabilities and governmental attempts to control them illustrate the interactions between ecosystems that influence the deployment of such systems. A-HMT-S can trigger social effects across multiple domains, including the labour market and political sphere, potentially leading to job creation or elimination, and regulatory crises. Particularly when it comes to AWS, these require updates to military training programs to ensure personnel are prepared to work with A-HMT-S. Additionally, an ecosystemic perspective reveals how these systems interact with their environments, creating new tasks and challenges. The deployment of A-HMT-S can significantly impact human activities, emphasising the need for explainable systems to address economic inequalities and build trust (Akiyoshi, 2022; Mayer, 2023).

Despite extensive research in these areas, there remains a specific knowledge gap in systematically analysing the risks and corresponding control measures in AWS development for military operations. This research aims to fill this gap by providing a critical reflection of these risks and proposing comprehensive control measures to mitigate them.

3. Research Methodology

To achieve the goal of this research, a systematic literature review approach is considered in alignment with the transparent PRISMA methodology (Denyer and Tranfield, 2009; Page *et al.*, 2021). This process begins with the establishment of well-defined objectives that outline the specific risks and corresponding control measures for tackling them. Following this, the identification of relevant studies is conducted using the IEEE Xplore, ACM, Taylor & Francis, SAGE Journals, and Wiley. Various combinations of keywords like AWS, Autonomous Weapon Systems, trustworthy, trustworthiness, military, and defence, ae used. Once the broad set of studies is collected (n = 10354), duplicates are removed, and a meticulous selection process is carried out by applying a series of inclusion and exclusion criteria such as the AWS and trust elements need to be included, written in English language, and they should be published between 01.01.2023 up to 31.12.2023. Further, the final set of articles is considered for an in-depth analysis. Among others, this stage implies identifying patterns, evaluating control measures, and understanding how risks are being addressed across different military contexts. At the end, the results obtained are described based on the process carried out, as depicted in Figure 1.

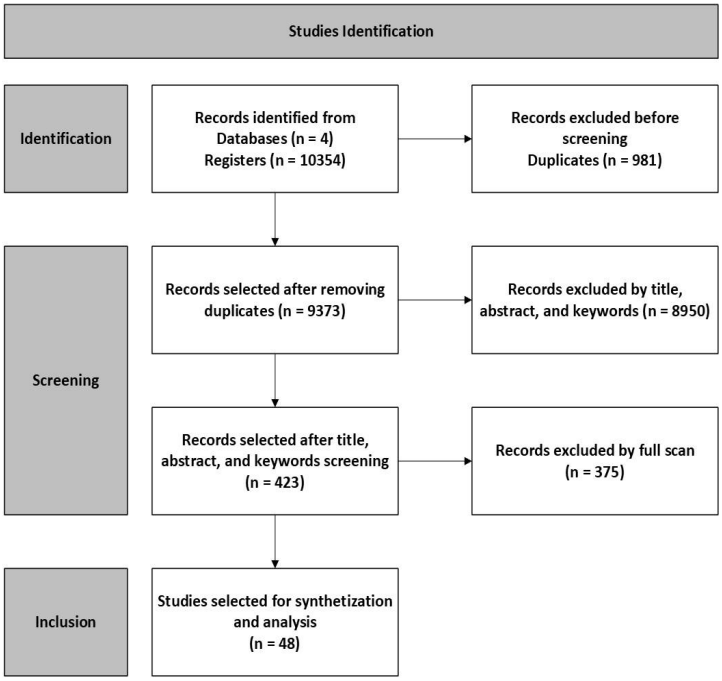

Figure 1: Methodological approach followed

4. Risks

Building AWS presents various risks that span technical and socio-technical dimensions, which must be carefully considered to ensure their trustworthy development and deployment. On the technical side, AWS face challenges related to system reliability, AI decision-making accuracy, and cyber security vulnerabilities.

These technical risks directly affect operational effectiveness and safety. At the same time, the socio-technical risks are characterised by ethical, social, and legal dimensions, and are equally important since they involve human-machine trust, ethical considerations, and accountability gaps.

To build reliable AWS, rigorous testing protocols must be applied, encompassing not only predictable, controlled scenarios but also complex, uncertain environments that mirror real-world combat conditions. This comprehensive testing mechanism helps to ensure that AWS can function as intended without causing unintended harm, even in dynamic and high-risk situations. Furthermore, legal reviews and validation processes are essential to confirm that AWS comply with ethical and legal standards, thereby mitigating the risk of unlawful engagements (Roff and Danks, 2018). At the same time, AI-decision making errors may occur when incorrect decisions lead to misidentifications in combat environments. Such errors can result in unlawful engagements and unintended harm to civilians as collateral damage, posing serious ethical and operational concerns (Galliott and Wyatt, 2022). Furthermore, the development process of AWS needs to be precise since issues such as the incorrect identification of target profile increases the likelihood of unintended engagements and minimises collateral damage (Bode, 2023). Moreover, if an AWS fails to function as expected or delivers unreliable outcomes, it risks eroding operator confidence, leading to operational failures and compromised missions (Albayram *et al.*, 2020). Trustworthy AWS must address the technical failures currently affecting drones, ensuring that they are dependable in the field (Zhai and Ye, 2020). For instance, target identification issues that UAVs could pose can result in collateral damage and civilian casualties, undermining both operational effectiveness and ethical standards. To build trust in AWS, it is crucial to enhance the precision of these systems and reduce targeting errors. This requires the implementation of advanced AI algorithms that are capable of accurately distinguishing between legitimate military targets and non-combatants. Then these models need to be properly tested and evaluated to ensure the reliability of these targeting systems in complex combat environments (Johansson, 2018). Automation bias presents significant risks when human operators either over-rely on the system's decisions or under-trust its capabilities, both of which can lead to critical operational failures. Over-trusting AWS may cause operators to disengage as they assume the system's decisions are always correct, while under-trusting may result in hesitation or ignoring valuable system input (Firlej and Taeihagh, 2021). Additionally, as AWS systems operate at high speeds, human operators are often assigned to supervisory roles, limiting their ability to actively control the system's decisions in real time. However, maintaining human oversight is essential for ensuring ethical judgement, especially in unpredictable or morally complex scenarios (Bode and Huelss, 2023). As AWS frequently operates at speeds that surpass human reaction time, risks occur when the human operators are unable to intervene in time to prevent potentially harmful decisions. Accordingly, mechanisms must be in place to ensure that operators have sufficient time to assess the system's decisions and intervene when necessary, particularly in critical or morally complex situations (Gubrud, 2014).

The proliferation of AWS to non-state actors and hostile regimes poses significant security risks, as these technologies could be misused in ways that destabilise global security. To build trust in AWS from a policy and governance perspective, it is essential to implement mechanisms that prevent the uncontrolled spread of this technology (Horowitz, 2021). Furthermore, the security of AWS communication systems is fundamental for maintaining operational integrity and effectiveness. UAVs need to employ robust encryption protocols and secure authentication methods to protect sensitive data and prevent threats such as unauthorized access or hijacking (Huang *et al.*, 2023). Additionally, AWS are vulnerable to cyber-attacks, such as jamming or impersonation, which could disrupt their functionality or allow adversaries to manipulate the system (Akter et al., 2023). In the context of coordination of multiple AWS across teams, particular challenges occur when missions are conducted. On this behalf, effective coordination requires sophisticated capabilities that ensure AWS operate in harmony with other systems and human operators, minimising the risk of miscommunication and operational inefficiencies that could jeopardise mission success. Then, trustworthy AWS must be designed to support real-time communication and coordination between teams, preventing potential breakdowns in system interaction that could compromise the mission's objectives (Stensrud, Valaker and Haugen, 2020). Additionally, while fully autonomous systems offer operational advantages, they also introduce unpredictability, particularly in situations where human oversight is insufficient. To mitigate the risks of autonomous decision-making, it is crucial to maintain "man-in-the-loop"/"man-on-the-loop" control structure (Brown, 2023). Specifically, AWS are highly susceptible to malware, which poses significant risks to their operational integrity and mission success. By safeguarding AWS from such attacks, their operational safety and system integrity can be preserved, ensuring that missions are not compromised and that AWS remain reliable under adversarial conditions (Théron and Kott, 2019).

Trust in human-machine collaboration is essential for the successful deployment of AWS, especially in dynamic and unpredictable combat conditions. For AWS to be deemed trustworthy, human operators should have confidence that the system will behave transparently, predictably, and respond appropriately to their input (Maathuis, 2023). This trust can be cultivated by ensuring that AWS behaviour aligns with operator expectations and that any autonomous actions are understandable and controllable (Roff and Danks, 2018). Moreover, a balance between AI decision-making autonomy and human oversight is necessary. Trustworthy AWS must be designed to augment, rather than supplant, human decision-making, ensuring that human ethical values and strategic considerations remain integral to the decision-making process (Mayer, 2023).

Ethical and moral concerns arise when building AWS as the human exposure to combat is reduced and the detachment of humans from lethal decision-making could directly impact the direct adherence to ethical warfare principles. Trustworthy AWS must incorporate mechanisms that allow human operators to remain engaged in ethical decision-making, either through direct oversight or simulated moral training, ensuring that human ethical responsibilities are not diminished or "deskilled" in combat scenarios (Cappuccio, Galliott and Alnajjar, 2022). Additionally, the use of AWS could imply various psychological implications and consequences on both soldiers and civilians, especially when dealing with collateral damage on civilians and civilian infrastructure (Wagner, 2014). Moreover, soldiers tasked with overseeing AWS may experience cognitive overload or discomfort from delegating life-or-death decisions to machines, while civilians may feel vulnerable and dehumanised under AWS surveillance and targeting (Wasilow and Thorpe, 2019).

While AWS can enhance operational efficiency by enabling faster decision-making and reducing human workload, their autonomy also demands that AWS remain adaptable and accountable to avoid strategic misalignment. This shift in mission structures requires that AWS be seamlessly integrated with human leadership to ensure that critical decisions remain under human control. Trustworthy AWS need to operate within established strategic frameworks, providing human commanders with the ultimate authority over key decisions. By ensuring that autonomy complements rather than replaces human leadership, AWS can avoid unintended outcomes and maintain alignment with broader military objectives (Mancini *et al.*, 2019). Nevertheless, building and maintaining calibrated trust between human operators and AWS is essential to ensuring their effective deployment. Over-trusting AWS can lead to excessive reliance on the system's capabilities, which may result in operators disengaging and missing critical opportunities for intervention, especially in complex scenarios. Conversely, under-trusting AWS may lead to underutilization, limiting the system's operational benefits (Bossuyt *et al.*, 2023). Then, it is essential that humans remain actively engaged, especially in morally and ethically complex situations where human judgement is crucial. Additionally, the decision-making pace of AWS should allow for meaningful human input, preventing the system from acting too quickly for human operators to respond (Bode and Huelss, 2023).

The AWS proliferation of AWS also introduces socio-technical risks regarding accountability. Trustworthy AWS must be equipped with robust security measures to prevent unauthorised use, ensuring that these systems are not exploited by groups who may operate outside the bounds of international law or ethical norms. As AWS become more widespread, particularly in asymmetrical warfare contexts, it is crucial to establish accountability mechanisms that can trace the system's usage and ensure compliance with legal and ethical standards (Haner and Garcia, 2019). In relation to this, robust regulation and governance becomes necessary since trustworthy AWS need to be developed, deployed, and used within strong regulatory frameworks that ensure compliance with international laws of war and ethical guidelines, preventing their misuse in both military and civilian contexts. These frameworks must clearly define the rules of engagement, ensuring that AWS are used in ways that align with global security standards and ethical warfare principles (Borson and Xu, 2022). Moreover, AWS introduce significant complexities in determining criminal and legal accountability, particularly when these systems are involved in unlawful actions. To mitigate the risks of accountability gaps, trustworthy AWS must include features that enable the tracking and auditing of decisions made by the system. Such mechanisms ensure that responsibility can be accurately assigned, whether to human operators, developers, or commanders, thereby maintaining a clear chain of command even in cases where AWS are responsible for lethal actions (Vallor, 2013). Furthermore, as AWS autonomy increases, concerns about a "responsibility gap" arise, where it becomes unclear who is accountable for the system's actions if human oversight is lacking (Schulzke, 2013).

In real-world environments, AWS raise critical human safety concerns, both for operators and civilians. To this end, they need to prioritise safety by incorporating robust failsafe mechanisms designed to prevent unintended harm, particularly in situations involving lethal engagements. These failsafe features are essential to mitigating risks associated with machine errors or unforeseen operational conditions that could otherwise

result in loss of life. Safety protocols must be rigorously tested and validated to ensure that human lives are not endangered due to technical malfunctions or unpredictable system behaviours (Boyles, 2021).

5. Control Measures

Following the identification of various risks associated with the development and deployment of AWS, this research examines these risks through a comprehensive framework of control measures. Several categories of control measures can be identified: Human-machine trust (Roff and Danks, 2018b; Gebru *et al.*, 2022), ethical and accountability frameworks (Brownsword, 2017; Gebru *et al.*, 2022), transparency and explainability (Warren and Hillas, 2020b), reliability and testing (Żmuda, 2023), design and development approaches (Warren and Hillas, 2020b; Gebru *et al.*, 2022), contextual and situational awareness (Gardner, 2021), regulation and governance (Warren and Hillas, 2020), and (cyber)security measures (Wasilow and Thorpe, 2019).

To begin with, human-machine trust is crucial for earning human trust in AI-driven autonomous systems, as explored by (Gebru *et al.*, 2022). These systems must overcome issues from (component) failures and opaque models. Trust calibration demands clear communication, understanding trust factors, and practical ethical frameworks through interdisciplinary collaboration. A simplistic "trust" framework is inadequate for AWS. Warfighters must grasp AWS operations with nuance beyond mere predictability. This requires comprehensive socio-technical adjustments, practical ethical frameworks, enhanced transparency, and trust calibration methodologies. The current military structure must evolve to support adaptive AWS trustworthiness (Roff and Danks, 2018).

Furthermore, the Defense Advanced Research Projects Agency addresses transparency and explainability through the Explainable Artificial Intelligence program. This program aims to create explainable models with high performance, helping users understand and trust AI systems. Recommendations include joint training, avoiding human-like system features to prevent over-confidence, and providing situational awareness context. Similarly, the Air Force's "Autonomous Horizons" report also highlights the importance of situational awareness as well as cognitive congruence, transparency, and human-system integration from the very beginning of the design process. Trust in AWS relies on these socio-technical adjustments, ensuring commanders can make informed decisions (Warren and Hillas, 2020).

In addition to these, ethical and accountability frameworks are essential in managing the deployment of AWS, as discussed in (Gebru *et al.*, 2022) and (Brownsword, 2017). Trustworthiness in AI-driven autonomous systems is essential for earning human trust and is often compromised due to failures and the opaque nature of "black box" models. Action-oriented frameworks and guidelines are necessary to translate ethical principles into practice, involving inputs from various disciplines and stakeholder consensus. A key concern is that technological management may interfere with agents' ability to act morally and freely. For instance, relying on machine learning for decision-making in life-and-death scenarios could erode human moral reasoning and dignity. It's argued that humans, not machines, should make such critical decisions to prevent a dehumanising effect. These frameworks must ensure that technological advancements uphold human oversight in the deployment of AWS. Proper calibration of trust between humans and machines relies on clear communication of operations, understanding trust factors, and aligning ethical standards with practical implementations to cultivate trustworthiness and acceptance in society.

Moreover, reliability and testing are critical for AWS due to operational failures, as highlighted in (Żmuda, 2023). These failures underscore the need for rigorous testing and validation to ensure AWS can perform reliably under various conditions. Testing dynamics involve a continuous process of identifying and addressing the limitations of these technologies through extensive trials and real-world simulations. This includes stress-testing AWS to evaluate their response to various (adversarial) scenarios and ensuring that they can function effectively within the complex technological networks they operate in.

AWS often struggle with acquiring and communicating contextually relevant information. Their ability to recognize objects in cluttered environments and adapt to new tasks is limited, exposing their "brittle" nature. Context is crucial for understanding the causes and consequences of actions, a nuance often missed by AI systems. Challenges such as informational uncertainties, adversarial deception, and the inherent opacity of machine learning further hinder AWS effectiveness. To overcome these issues, AWS design must emphasise contextual awareness. Enhancing AI capabilities to ensure accurate, adaptive responses, and maintaining transparency in decision-making processes are vital to mitigate uncertainties and build trust (Gardner, 2021).

Furthermore, regulation and governance are critical for the effective deployment of AWS, as highlighted by Warren & HIllas (2020). Effective regulation and governance must incorporate principles such as cognitive

congruence, transparency, situational awareness, human-systems integration, and joint training from the design process's outset to maximise trust in machine-assisted decision-making. Ensuring human control and oversight in AWS operations is essential, as military personnel must make informed decisions rather than merely authorising actions proposed by machines. This requires evolving military structures and policies to support the integration and trustworthiness of adaptive AWS, ensuring they comply with international humanitarian law and public acceptance.

Lastly, cybersecurity and security measures are essential for AWS, as highlighted in (Wasilow and Thorpe, 2019). AI researchers face challenges with reproducibility and the risk of malicious use of open-source code. Ensuring robust validation and verification is essential to address potential operational risks, such as malfunctions, adversarial interference, and unexpected behaviours. AI technologies must include comprehensive cybersecurity protocols to protect against algorithmic counterfeits, adversarial attacks, and other vulnerabilities. Effective security measures are vital to safeguard the integrity and reliability of AWS, ensuring these systems function as intended in various operational contexts and comply with ethical and legal standards.

6. Conclusions

Building and deploying AWS is important for both ensuring operational effectiveness as well as to maintain compliance with the relevant socio-technical dimensions that include compliance with international laws, ethical standards, and social perspectives. This means that trustworthy AWS need to be designed in a safe, responsible, and reliable manner to properly function in unpredictable combat events, uphold ethical decision-making principles, and incorporate robust control mechanisms to prevent unintended effects such as collateral damage on civilian and civilian infrastructure. The trust strategic and military leadership, policy makers, and operators place in these systems directly impacts the success of military missions and the broader acceptance of AWS in military operations. Additionally, ensuring transparency, accountability, and human involvement at critical decision points is essential to mitigating the risks associated with autonomous decision-making, thereby preserving legal and ethical standards and principles in warfare.

To this end, this research provides valuable insights into the key risks and control measures associated with building trustworthy AWS through a systematic literature review. The findings show a wide range of technical risks, such as AI decision-making errors, system reliability issues, and cyber security vulnerabilities, as well as socio-technical risks involving human-machine trust dynamics, ethical accountability, and the challenges of governance and regulation. To mitigate these risks, the literature emphasises the need for advanced fail-safes, rigorous testing and validation, and strong regulatory frameworks. Further, two promising research future research areas emerge from this study. First, further investigation into human-AI teaming trust calibration techniques (Maathuis, 2024) that have the potential to enhance our understanding of how to maintain balanced trust between operators and AWS, ensuring neither over-reliance nor underutilization of the systems. And second, developing improved responsibility and accountability mechanisms that trace decision-making processes in AWS from design to deployment for addressing the growing "responsibility gap" and ensuring that ethical and legal responsibility is clearly assignable in the context of increasingly autonomous systems. Adopting critical lenses in these areas further contribute to advancing the responsible, safe, and trustworthy development of AWS.

References

Akiyoshi, M. (2022) 'Trust in things: A review of social science perspectives on autonomous human-machine-team systems and systemic interdependence', *Frontiers in Physics*, 10, pp. 951296.

Akter, R., Golam, M., Doan, V. S., Lee, J. M., & Kim, D. S. (2022). Iomt-net: Blockchain-integrated unauthorized uav localization using lightweight convolution neural network for internet of military things. *IEEE Internet of Things Journal*, *10*(8), 6634-6651.

Albayram, Y. *et al.* (2020) 'Investigating the effects of (empty) promises on human-automation interaction and trust repair', in *Proceedings of the 8th international conference on human-agent interaction*, pp. 6–14.

Bakx, G.C.H. and Nyce, J.M. (2017) 'Risk and safety in large-scale socio-technological (military) systems: a literature review', *Journal of Risk Research*, 20(4), pp. 463–481. Available at: https://doi.org/10.1080/13669877.2015.1071867.

Bode, I. (2023) 'Practice-based and public-deliberative normativity: retaining human control over the use of force', *European Journal of International Relations*, 29(4), pp. 990–1016.

Bode, I. and Huelss, H. (2023) 'Constructing expertise: the front- and back-door regulation of AI's military applications in the European Union', *Journal of European Public Policy*, 30(7), pp. 1230–1254. Available at: https://doi.org/10.1080/13501763.2023.2174169.

Borson, J.E. and Xu, H. (2022) 'A Path Dependent Approach for Characterizing the Legal Governance of Autonomous Systems', *IEEE Access*, 10, pp. 119985–119998.
Bossuyt, D.L. Van *et al.* (2023) 'Trust Loss Effects Analysis Method for Zero Trust Assessment', in *2023 Annual Reliability and Maintainability Symposium (RAMS)*, pp. 1–6.
Boyles, R.J.M. (2021) 'Hume's law as another philosophical problem for autonomous weapons systems', *Journal of Military Ethics*, 20(2), pp. 113–128.
Brown, A. (2023) 'Ethics, autonomy, and killer drones: Can machines do right?', *Comparative Strategy*, 42(6), pp. 731–746.
Brownsword, R. (2017) 'From Erewhon to AlphaGo: for the sake of human dignity, should we destroy the machines?', *Law, Innovation and Technology*, 9(1), pp. 117–153.
Cappuccio, M.L., Galliott, J.C. and Alnajjar, F.S. (2022) 'A taste of armageddon: A virtue ethics perspective on autonomous weapons and moral injury', *Journal of Military Ethics*, 21(1), pp. 19–38.
Cools, K. and Maathuis, C. (2024) 'Trust or Bust: Ensuring Trustworthiness in Autonomous Weapon Systems', *arXiv preprint arXiv:2410.10284* [Preprint].
Dekker, S.W.A. (2011) 'Drift into Failure: From Hunting Broken Components to Understanding Complex Systems', in. Available at: https://api.semanticscholar.org/CorpusID:61256182.
Denyer, D. and Tranfield, D. (2009) 'Producing a systematic review.'
Firlej, M. and Taeihagh, A. (2021) 'Regulating human control over autonomous systems', *Regulation & Governance*, 15(4), pp. 1071–1091.
Galliott, J. and Wyatt, A. (2022) 'Considering the importance of autonomous weapon system design factors to future military leaders', *Australian Journal of International Affairs*, 76(2), pp. 219–244.
Gardner, N. (2021) 'Clausewitzian friction and autonomous weapon systems', *Comparative Strategy*, 40(1), pp. 86–98.
Gebru, B. *et al.* (2022) 'A review on human–machine trust evaluation: Human-centric and machine-centric perspectives', *IEEE Transactions on Human-Machine Systems*, 52(5), pp. 952–962.
Gubrud, M. (2014) 'Stopping killer robots', *Bulletin of the Atomic Scientists*, 70(1), pp. 32–42.
Haner, J. and Garcia, D. (2019) 'The artificial intelligence arms race: Trends and world leaders in autonomous weapons development', *Global Policy*, 10(3), pp. 331–337.
Horowitz, M.C. (2021) 'When speed kills: Lethal autonomous weapon systems, deterrence and stability', in *Emerging technologies and international stability*. Routledge, pp. 144–168.
Huang, Y. *et al.* (2023) 'A Review of Authentication Methods in Internet of Drones', in *2023 International Conference on Networking and Network Applications (NaNA)*, pp. 7–12.
Johansson, L. (2018) 'Ethical aspects of military maritime and aerial autonomous systems', *Journal of Military Ethics*, 17(2–3), pp. 140–155.
Johnson, J. (2022) 'The AI commander problem: Ethical, political, and psychological dilemmas of human-machine interactions in AI-enabled warfare', *Journal of Military Ethics*, 21(3–4), pp. 246–271.
Longpre, S., Storm, M. and Shah, R. (2022) 'Lethal autonomous weapons systems & artificial intelligence: Trends, challenges, and policies', *Edited by Kevin McDermott. MIT Science Policy Review*, 3, pp. 47–56.
Maathuis, C. (2023) 'Human Centered Explainable AI Framework for Military Cyber Operations', in *MILCOM 2023-2023 IEEE Military Communications Conference (MILCOM)* (pp. 260-267). IEEE.
Maathuis, C. (2024) 'Trustworthy Human-Autonomy Teaming for Proportionality Assessment in Military Operations', in *2024 4th International Conference on Applied Artificial Intelligence (ICAPAI)*, pp. 1–8.
Mancini, F. *et al.* (2019) 'Securing Autonomous and Unmanned Vehicles for Mission Assurance', in *2019 International Conference on Military Communications and Information Systems (ICMCIS)*, pp. 1–8.
Mayer, M. (2023) 'Trusting machine intelligence: artificial intelligence and human-autonomy teaming in military operations', *Defense & Security Analysis*, 39(4), pp. 521–538.
Page, M.J. *et al.* (2021) 'Declaración PRISMA 2020: una gu\'\ia actualizada para la publicación de revisiones sistemáticas', *Revista española de cardiolog\'\ia*, 74(9), pp. 790–799.
Pedron, S.M. and da Cruz, J. de A. (2020) 'The future of wars: Artificial intelligence (ai) and lethal autonomous weapon systems (laws)', *International Journal of Security Studies*, 2(1), p. 2.
Roff, H.M. and Danks, D. (2018) '"Trust but Verify": The difficulty of trusting autonomous weapons systems', *Journal of Military Ethics*, 17(1), pp. 2–20.
Rosendorf, O., Smetana, M. and Vranka, M. (2024) 'Algorithmic Aversion? Experimental Evidence on the Elasticity of Public Attitudes to "Killer Robots"', *Security Studies*, 33(1), pp. 115–145.
Schraagen, J.M. (2023) 'Responsible use of AI in military systems: prospects and challenges', *Ergonomics*, 66(11), pp. 1719–1729. Available at: https://doi.org/10.1080/00140139.2023.2278394.
Schulzke, M. (2013) 'Autonomous weapons and distributed responsibility', *Philosophy & Technology*, 26, pp. 203–219.
Stensrud, R., Valaker, S. and Haugen, T. (2020) 'Interdependence as an Element of the Design of a Federated Work Process', in *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pp. 1–6.
Théron, P. and Kott, A. (2019) 'When autonomous intelligent goodware will fight autonomous intelligent malware: A possible future of cyber defense', in *MILCOM 2019-2019 IEEE Military Communications Conference (MILCOM)*, pp. 1–7.
Vallor, S. (2013) 'The future of military virtue: Autonomous systems and the moral deskilling of the military', in *2013 5th International Conference on Cyber Conflict (CYCON 2013)*, pp. 1–15.

Vaughan, D. (1996) *The Challenger launch decision: Risky technology, culture, and deviance at NASA*. University of Chicago press.

Wagner, M. (2014) 'The dehumanization of international humanitarian law: legal, ethical, and political implications of autonomous weapon systems', *Vand. J. Transnat'l L.*, 47, p. 1371.

Warren, A. and Hillas, A. (2020) 'Friend or frenemy? The role of trust in human-machine teaming and lethal autonomous weapons systems', in *Robotics, Autonomous Systems and Contemporary International Security*. Routledge, pp. 132–160.

Wasilow, S. and Thorpe, J.B. (2019) 'Artificial intelligence, robotics, ethics, and the military: A Canadian perspective', *AI Magazine*, 40(1), pp. 37–48.

Zajac, M. (2023) 'Aws Compliance with the Ethical Principle of Proportionality: Three Possible Solutions', *Ethics and Information Technology*, 25(1), pp. 1–13. Available at: https://doi.org/10.1007/s10676-023-09689-8.

Zhai, Q. and Ye, Z.-S. (2020) 'How reliable should military UAVs be?', *IISE Transactions*, 52(11), pp. 1234–1245.

Żmuda, M.D. (2023) 'Autonomous weapons of pleasure. Media archaeology of automated killing in military and gaming technologies', *Culture, Theory and Critique*, pp. 1–21.